Journal of Computer Science

Original Research Paper

A Systematic Literature Review for Implementing Data Ops in the Data Warehouse Lifecycle during the ETL Phase

[1]Ahmed Bahaa, [2]Sherif Ragab Eldemerdash and [3]Hanan Fahmy

[1]Faculty of Computers and AI Helwan University, Faculty of Computers and AI Beni-suef University Egypt
[2,3]Faculty of Computers and AI Helwan University Egypt

Article history
Received: 23-06-2021
Revised: 27-09-2021
Accepted: 08-10-2021

Corresponding Author:
Sherif Ragab Eldemerdash
Information Systems Dep.,
Faculty of Computer and AI,
Helwan University, Egypt, Egypt
Email: ahmed.bahaa@fci.helwan.edu.eg,

Abstract: Nowadays, no one can deny the importance of Data Ware House (DWH) in all organizations. The most important components in Data Ware House (DWH) are the Extraction, Transformation, Loading (ETL) phase. Data cleaning is a basic piece of the transformation stage in Data Warehousing. This may affect critical activities such as data collection and decision-making in various organizations Data Ops is an evaluation technique of Dev Ops in the data domain. This study conducts a Systematic Literature Review (SLR) to assess the previous studies of data warehouses related to Data Ops efforts. This study collects 55 primary studies related to the detection of Data Scrubbing, Data Consistency, Data warehouse, Dev Ops and Data Ops and we have conducted a Systematic Literature Review (SLR). Based on these findings, we discuss many concerns related to the study of current approaches in terms of abstraction level, metrics used, implementation and validation. That is why the analysis covers the published efforts between 2016 and 2021 since Data Ops is a significantly new technique. The survey should cover only research that took plan in recent years. The result of the study observed that 29% of the studies focused on solving the importance of data quality in the data warehouse, 62% of them focused on related Dev Ops, only 9% focused on Data Ops techniques and no 0% survey on enhancing ETL phase with Data Ops. This SLR brings to the attention of the research community several opportunities for using Data Ops in future research and the nearly proposed model DW Ops.

Keywords: Data Scrubbing, Data Quality, Data Ware House (DWH), Dev Ops, Data Ops, ETL, Data Transformation

Introduction

A data warehouse is a home for your high-esteem data or knowledge tools that start with different business applications. Data warehousing is the planned, architected, coordinated and intermittent replicating of data from numerous sources, both inside and outside the endeavor, into a domain enhanced for analytical and informational processing. Data Warehousing includes encouraging change in business forms. In expansion to improving data-driven operational and strategic choices, organizations gain knowledge into key zones that Code can assist organizations with making vital choices about the principal parts of their business (Hammergren, 2009).

Data Warehouse supports many heterogeneous sources of data with different characteristics for quality, speed and structure. Data Warehouse is used to help Decision-makers by analyzing those data using OLAP tools. It empowers administrators to get to the data they need in a practical way for settling on the correct choice for any work. To be able to make the correct decision, the analysis tools must be of excellent quality, so as not to create a wrong decision. The powerful Data Quality (DQ) is an important factor for a better decision.

Data quality has been commonly used in far too many areas, such as health, banking, corporate organization and the information system. It included the need to particularly comprehend the quality of the elements of data, measurement methods, assessment techniques and improvement processes in each area. Data Quality and the effect of having low data quality are still in conversations and research (Izham Jaya, 2019). A large amount of data are loaded daily in the data warehouse from different sources makes Data Quality a most critical challenge in almost all forms of data analytics. Data quality is sensitive and critical for the achievement of Data analysis. The data stored in the data warehouse must be right, perfect and high caliber. High data quality in the data warehouse will

bring about better analysis and decision-making processes. So, this data quality issue must be dealt with before the data is stacked into the data warehouse (Azeroual *et al*., 2019a). Poor data quality caused a disaster in the decision, So, ETL use tools to make data scrubbing.

Data scrubbing is also called data cleansing. Scrubbing of Data is the method of removing or improving dirty data in the ETL phase. Data cleansing is erroneous, formatted improperly, incomplete, or various. Dirty records in the data source mean incomplete, incorrect, redundant, redundant records, or out-of-date data. A big company used the data warehouse (insurance, banking, telecommunications, or transportation, etc.). This company might use a data cleansing use rules or algorithms and search tables to quickly look at data flaws and get efficient analytics (Rainardi, 2008). Data scrubbing remains expensive and time-consuming in data quality and affects the effort of analysts. ETL phase needs a developer to write data scrubbing rules and algorithms manually. They execute as long batch jobs. The tools help analysts define rules, automatically data quality and correction to optimize these rules (Michele Dallachiesa, 2013). Therefore, we need a tool to assist developers in the work and enhance data scrubbing in ETL at the same time.

Dev Ops (Development and Operation) is a venture software improvement that utilizes agile and a pooled connection between its operation and developer. Dev Ops has for the past few years become a fundamental part of software housing. The Dev Ops objective is to abbreviate the life cycle of the development system while likewise conveying highlights, fixes and refreshes every now and again in close arrangement with business objectives (Vinicius Lima Cruz, 2018). Dev Ops is an advancement to agile development from Continuous Development (CD), continuous delivery, Continuous Integration (CI), continuous testing, continuous deployment and continuous monitoring. Dev Ops centers on the automation of change, configuration and release processes (Saima Rafi, 2020).

Today, in an environment in which competitive advantage demands fast time to market and unceasing experimentation. The organizations that use Dev Ops can have many modified practices many or hundreds of times per day. Organizations that do not reproduce these findings are likely to lose more agile rivals in the commercial center and may altogether quit the industry, just like the assembling associations that did not follow the Lean concept. The Software Housing enables developers safely to autonomously develop, test and deploy applications to the clients. Dev Ops has made a revolution in a software house. It had to speed up the development, test, deployment and monitor. The most recent Software Development and version took a large number of dollars and quit while developing, deploying and releasing. When dealing with data coming from heterogeneous data, Dev Ops still faces problems.

Dev Ops takes into account the question of the data quality assessment process. The topic of the approach to data quality in the Dev Ops methodology was not discussed. The difficulties indicate the weak zones that should be tended to for the achievement and progression of software projects (Saima Rafi, 2020). Therefore, software houses and researchers try to find a new technique to help speed up release output and analytic teams. They found a new approach deal with data. They named it Data Ops. They need Data Ops to support companies with more effective data analytics to make better decisions through more effective data analytics.

Related Work

In what follows, we will include a description of the enhanced data quality using Data Ops specific to our research, including current SLRs, in what follows. Not only will this review allow up-to-date coverage of literature, but it will also illustrate the merits of our proposed survey.

Prabin and Anshu proposed a survey on the concept of Data Ops and how its adoption across industries is gaining momentum. They contrasted the notions of Data Ops and Dev Ops. Then it reflects on the value of Data Ops in the industrial and service sectors. They described the mechanism and platform of Data Ops as well as the data issues in the sectors of manufacturing & utilities. Various Data Ops techniques are also addressed for these sectors, alongside the significance of implementing applied analytics through Data Ops to achieve market benefits (Sahoo and Premchand, 2019).

Saima, Wu, Muhammad and Ahmed suggested a survey on the crucial variables that may have a detrimental effect on the method of assessing data quality in Dev Ops. They used the Systematic Literature Review (SLR) method and established a total of 13 critical problems. A questionnaire survey of industry experts further checks SLR findings. They used the Blurry TOPSIS method to prioritize the daunting variables analyzed concerning their relevance in the measurement phase of Dev Ops data quality. The findings reveal that real-time data collection, data visualization, incomplete data and other invalid data are the top-ranking issues that need to be solved on a priority basis to efficiently assess the output of heterogeneous data in Dev Ops (Saima Rafi, 2020).

Aiswarya, Jan Bosc and Helena were driven by direct a direct qualitative multiple-case study and interviews with the representatives of three companies. They identified the key challenges and benefits associated with data pipeline implementation and use. With five use cases from three case companies, based on multiple case study research. They were clear about the importance of implementing data pipelines to allow traceability, fault tolerance and reduce human errors by maximizing automation to produce high-quality data (Munappy *et al*., 2020).

Damian, Willem-Jan and Martin attempt to tackle insight issues with regards to a work market utilizing

Information investigation upheld by artificial intelligence calculations to empower abilities limitation and recovery. Via unique Data Ops models, have been formulating and solving this problem, combining data sources from administrative and technological partners in many countries into collaboration, the building required expertise to policy and decision-making support. They focused on the indispensable occupation of eliminating mastery from resumes and openings highlighting cutting-edge machine learning models (Tamburri *et al*., 2020).

Our survey identified 55 research papers and SLR that focused on using Data Ops to improve the quality of data. These papers included sixteen research papers that focused on data scrubbing and data quality, with thirty-four research papers focusing on Dev Ops and its relationship to data quality improvement, in addition to. Five research papers that focused on Data Ops.

Based on the review outlined above, our SLR differs from existing ones as follows, we try to use Data Ops to improve the quality of data in the ETL phase in data warehousing as coming in the rest of the research parts.

Research Method

As suggested by Kitchenham, this investigation has been embraced as a Systematic Literature Review (SLR) given the first rules (Kitchenham, 2004). The present SLR provides an in-depth review and discussion of current state-of-the-art research in data quality, Dev Data Ops and Data Ops with data warehousing. The review aims to take papers about our studies, so this analysis is classified as a tertiary A Systematic Literature Review (SLR) survey. The steps are described below in the Systematic Literature Review (SLR) method. As adopted in our Systematic Literature Review (SLR), the details of these guidelines are described in the next section.

Investigation Questions

The proposed Systematic Literature Review (SLR) identifies first the existing literature on quality of data, Dev Data Ops and Data Ops with data warehousing. Then, it investigates and answers the IQs. By answering these RQs, our SLR explores and investigates the findings reported in each of the selected research papers. To achieve generality, we consider all published literature related to data quality, Dev Data Ops and Data Ops with data warehousing techniques at the ETL Phase. Therefore, this study has the questions of research as below:

- IQ1. How much operation and activity have been there since 2016 with Data Ops?
- IQ2. What surveys are being classified in Data Ops?
- IQ3. What research topics are being data scrubbing by Data Ops?
- IQ4. What are the limitations of our survey?

Inclusion and Exclusion Criteria

A Systematic Literature Review (SLR) guidelines suggested in (Kitchenham, 2004), to pick the relevant research paper studies and weed out the irrelevant ones.

Inclusion Criteria

To select a research paper or a Systematic Literature Review (SLR) study, it must satisfy the following inclusion criteria:

1. It must be published between 1st January 2016 and 31st May 2021 in a journal or conference in advance. Our search query is constrained to this date in all electronic databases to ensure effectiveness.
2. Its theme is mainly related to the documentation and detection quality of data, Dev Data Ops and Data Ops with data warehousing.
3. It must introduce one or more quality of data, Dev Data Ops and Data Ops with data warehousing techniques at either the design or code level or both
4. A systematic Literature Review (SLR) for example writing studies known as exploration questions, search measure, data extraction and data presentation. In the event that the researchers alluded to their investigation as A Systematic Literature Review (SLR) or not.
5. Meta-Analyses (MA).
6. Articles that discuss the steps used for data scrubbing, data quality, Dev Ops and Data Ops or together.
7. Unofficial literature surveys (no defined research questions; no defined search measure, no defined data extraction measure).
8. At the point when we discovered copy articles from similar examinations, ("when a few reports of an investigation exist in various diaries, the most complete rendition of the examination was remembered for the audit").

Note that we have included articles containing data cleansing, data quality, Dev Ops and Data Ops, or a combination of them, the only feature of the article was the paper as well as papers for which the main object of the article was data scrubbing, data quality, Dev Ops and Data Ops.

By IQ1, it might be a worry that we began our survey beginning of 2016. To answer IQ1, we explicit the quantity of a Systematic Literature Review (SLR) distributed every year. We determine the number of Systematic Literature Reviews (SLR) issued in the journal/conference and whether the EBSE papers are cited or not (Dyba *et al*., 2005; Kitchenham *et al*., 2004) or Guidelines paper (Kitchenham, 2004).

To answer IQ2, we see the extent of the checking (“i.e., regardless of whether it tended to an innovation-

focused research question or whether it saw research patterns") and the software engineering point area.

To answer IQ3, we considered individual researchers. The association of which researchers are affiliated and the nation where the association is based.

Concerning the limitations of Systematic Literature Reviews (SLR) (IQ4). We put some questions to answer a Literature Reviews (SLR) limitation:

- IQ4.1. Were the subject points restricted?
- IQ4.2. is there evidence that Data Ops has been applied to fix data scrubbing because of a lack of primary research?
- IQ4.3. Is the goodness of Data Ops sufficient, improve if not?
- IQ4.4. Are Data Ops factors elements to solve data scrubbing?

Exclusion Criteria

The research paper consisted of the following, filtered out using the inclusion criterion:

1. Articles where the quality of data, Dev Ops process, Dev Data Ops and Data Ops with data warehousing are not the main focus Articles that do not present a novel or existing technique for design or code smell detection
2. News, novels and monographs in non-English
3. Articles with incomplete texts
4. Articles that lack a detailed description of the quality of data, Dev Ops process, Dev Data Ops and Data Ops with data warehousing techniques such as demonstration articles.
5. Studies that appear to support non-duplicates in multiple electronic libraries

Process of Search

The request cycle was a manual search of specific conference procedures, journal papers between the range of 2016 and 2021 and internet links. The internet links of data used in this survey are shown in Table 1. The chosen conferences and journals are shown in Table 2. We chose journals that incorporate detailed examinations of literature surveys. They have been used as sources for another Systematic Literature Review (SLR) specific to our topic. The incorporation and exclusion criteria and rules utilized in this investigation are adjusted from (Kitchenham and Charters, 2007; Mariano *et al*., 2017):

a) Articles were published in the English language.
b) Reports about the data quality or the quality of knowledge.
c) Articles, which are capable to answer at least one of our research questions.

Each journal and conference procedure was checked on by one of four distinct researchers. The researcher is responsible for looking at the particular journal or conference utilized to the related papers the point-by-point integration and rejection models (see Section 3.5). At this point, any included and excluded papers were checked by another researcher.

Quality Evaluation

After fulfilling the inclusion and exclusion requirements, we conducted a content evaluation to determine the quality of the study papers collected. On assessment criteria, based on Kitchenham guidelines (Kitchenham, 2004), we found that during the selection process for the survey, the quality valuation of the chosen research study was aimed at determining the efficacy and viability of the selected studies. Therefore, we have a four-Quality Evaluation (QE). The instructions given by Chen (2018; Inayat *et al*., 2015) were followed up in the way of this (QE). The quality estimation of the chosen fulfillment study is aimed through the study choice process, to decide the performance and viability of the chosen studies Chen (2018; Inayat *et al*., 2015; Khan *et al*., 2011). The indicators depend on four questions relating to Quality assessment (QE):

- QE1. Are the survey's incorporation and prohibition measures portrayed and suited?
- QE2. Does the quest for literature aim to involve all related research?
- QE3. Did the reviewers assess the quality/legitimacy of the inquiries that were included?
- QE4. Were the critical data/studies defined adequately?

Table 1: Links used in this study

Digital databases link	"http://ieeexplore.ieee.org"
	"http://dl.acm.org"
	"http://link.springer.com"
	"http://www.wiley.com"
	"http://www.sciencedirect.com"
	"http://www.scholar.google.com"
	"https://academic.microsoft.com"
	"https://www.ekb.eg/ar/web/"
	"https://www.researchgate.net"
Searched items	Books chapter, Conferences, Journals and Workshop articles
Language	English

Table 2: Journals and conference

Source	Acronym
ACM/IEEE International Conference on Software Engineering	ICSE
ACM Computing Surveys	ACS
ACM SIGPLAN Conference on Programming Language Design and Implementation	PLDI
ACM SIGSOFT International Symposium on Foundations of Software Engineering	ISFSE
ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages	POPL
Advances in Engineering Software	AIES
Applied Soft Computing Journal	ASE
Automated Software Engineering	AUSE
Business and Information Systems Engineering	BISE
Empirical Software Engineering	ESE
Engineering with Computers	EWC
EUROPEAN JOURNAL OF INFORMATION SYSTEMS	EJIS
Foundations and Trends in Machine Learning	FTML
IEEE Technology & Engineering Management Conference	TEMSCON
IEEE International Conference on Information Reuse and Integration for Data Science	IRI
IEEE International Conference on Software Maintenance and Evolution	ICSME
IEEE Software	IEEE SW
IEEE Transactions on Software Engineering	TSE
IET Software	IETSW
Information and Software Technology	IST
Information Systems Frontiers	ISF
International Conference on Software Analysis, Evolution and Reengineering	SANER
International Journal of Computer Vision	IJCV
International Journal of Learning, Teaching and Educational Research	IJLTER
International Journal of Human Capital and Information Technology Professionals	IJHCIT
International Journal of Intelligent Systems	IJIS
International Journal of Recent Technology and Engineering	IJRTE
International Journal of Software Engineering and Knowledge Engineering	IJSEKE
Information Systems Journal	ISJ
International Symposium on Software Testing and Analysis	ISSTA
Journal of Digital Information Management	JDIM
Journal of Scientific Computing	JSC
Journal of Software: Evolution and Process	JSEP
Journal of Software: Practice and Experience	JSPE
Journal of Software Testing, Verification, Reliability	JSTVR
Journal of Systems and Software	JSS
Journal of Theoretical and Applied Information Technology	JTAIT
Program	PR
Requirements Engineering	RQE
Science of Computer Programming	SCP
Software and Systems Modeling	SSM
Software Quality Journal	SQJ
Transactions on Software Engineering and Methodology	TSEM
Wiley Software: Evolution and process	WSEP

We have scored the questions as follows:

- QE1: Y (yes), the inclusion measures in the study are described explicitly. P (Partly), the measures for inclusion are tacit. N (no), inclusion measures are not specified and cannot easily be inferred from them.
- QE2: Y, the authors either searched at least four or more computerized databases and implemented additional search techniques or listed all articles that tend to the subject of interest and referred to them as (P). With no additional search strategies, the authors looked through three or four computerized libraries or looked through a specified but restricted scheme of conference and journal procedures. The authors search N, up to 2 computerized database, or an extremely restricted collection of journals.
- QE3: Y, the authors have explicitly defined quality metrics and excluded them from any critical analysis. P, the research concern requires quality concerns that the research tackles. No detailed quality evaluation of individual primary studies was attempted.
- QE4: Y, the information on each analysis is given. P, only summary data on critical studies is discussed. N, the outcomes of the individual basic studies are not stated.

We were scoring as Y = 1, P = 0.5, N = 0.

3.6. Data Gathering

The information gathered from each research was obtained:

- The origin and comprehensive reference (conference or journals).
- Research category classification (SLR, Meta-Analysis (MA)). Scope (investigation Question of research patterns or relevant technology assessment).
- Main point area.
- The writers and their organization and the country in which it is based.
- Report overview, including the main study questions and the answers.
- Investigation Questions/problems.
- Quality assessment.
- Regardless of whether the examination referred to the data scrubbing, data quality, Dev Ops and Data Ops or a combination between them, papers (Dyba *et al*., 2005; Kitchenham *et al*., 2004) or the Systematic Literature Review (SLR) Guidelines (Kitchenham, 2004).
- In any case, of whether the inquiry is proposed, it's based on specialist rules.
- How many simple researches are used or mixed between data scrubbing, data quality, Dev Ops and Data Ops?

One scientist collected the data and another tested the extraction process. With the clinical concepts summarized, the method of having one extractor and one inspector is not effective in Kitchen-ham's rules (Kitchenham, 2004), however, it is a technique we had discovered helpful by and by (Brereton *et al*., 2007).

Analyzing Data

The details that were tabulated to present the following information:

- The number of articles that have been published and the source of those papers per year (classify IQ1).
- If the papers referenced the scrubbing of data, data quality, Dev Ops and Data Ops or a combination between them (classify IQ1).
- The total of studies in each article evaluates (classify IQ2 and IQ4.1).
- The investigated subject by the scrubbing of data, data quality, Dev Ops and Data Ops or combination between them (classify IQ2 and IQ4.1).
- The affiliations of writers and their institutions (classify IQ3).
- In each paper, the number of research articles (classify IQ4.2).
- For each paper, the quality score (classify IQ4.3).
- If the articles proposed practitioner-oriented advice, (classify IQ4.4).

Search Outcomes

The current survey attempts to provide a comprehensive account of existing data quality literature, Dev Data Ops and Data Ops, all with data warehousing. The aims of the survey are expressed in the IQs described in section 3. This section summarizes our survey results. We first presented our search result including the number of likely articles, relevant articles and selected articles according to article resources and evaluate the results of the quality evaluation of each research paper considered in our sample. Next, we discussed the overview of selected articles.

Search Outcomes

Table 3 presents the outcomes of the search process according to the sources of the article and the year is written. Fifty-five articles were recovered during the search procedure. Thirty-three articles were prospective and fifteen were relevant to this study. Then we surveyed the relevant articles and potentially related articles and finally, six paper was chosen to be included in this survey. Our result showed that IEEE and ACM/IEEE published most of the data quality journal articles. As we showed in Fig. 1. 67% of a research paper is prospective, 30% of a research paper is relevant and 3% is selected.

Quality Assessment of Papers

Using the Database of Abstracts of Reviews of Effects (DARE) criterion, we analyzed the studies for consistency (see Part 3.4). The ranking for each analysis is shown in Table 4.

The quality of consistency of the finding reveals that all studies on the DARE scale scored 1.5 or more and 32 studies scored less than 3.55 (Jørgensen, 2007; Zannier *et al*., 2006). One scored four (Jørgensen, 2004; Kitchenham *et al*., 2006).

Quality Factors

The overall research of quality scores for each year is shown in Table 5. Note, "for this analysis we used the first publication date for each duplicated study". Table 5 shows that the number of published papers every year was very constant. A growing number of papers tend to have an average quality score. Figure 2 shows the percentage of quality score of paper.

Table 3: Sources searched for the years 2016/2021

	Source	Year	Prospective	Relevant	Selected
1	SAI Computing Conference	2016	1	0	0
		2017	0	0	0
		2018	0	0	0
		2019	0	0	0
		2020	0	0	0
2	IEEE	2016	2	2	0
		2017	2	1	0
		2018	5	0	0
		2019	4	0	0
		2020	0	0	1
3	IEEE Technology and Engineering Management Conference	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	0	0	1
		2020	0	0	0
4	ACM/IEEE	2016	1	0	0
		2017	0	0	0
		2018	2	0	0
		2019	2	1	0
		2020	0	0	0
5	ACM	2016	1	1	0
		2017	1	0	0
		2018	0	0	0
		2019	0	1	0
		2020	0	0	0
6	CEUR-WS Journal (open source)	2016	0	0	0
		2017	0	0	0
		2018	1	0	0
		2019	0	0	0
		2020	0	0	0
7	International Journal of Applied Information Systems (IJAIS)	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	1	0	0
		2020	0	0	0
8	arXiv journal (open source)	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	1	2	0
		2020	0	0	0
9	Journal of Software: Evolution and Process	2016	0	0	0
		2017	1	0	0
		2018	0	0	0
		2019	0	0	0
		2020	0	0	0
10	International Journal of Advanced Research in Computer Science	2016	0	0	0
		2017	1	0	0
		2018	0	0	0

Table 3: Continue

		2019	0	0	0
		2020	0	0	0
11	International Journal of Computer Applications	2016	0	0	0
		2017	0	0	0
		2018	1	0	0
		2019	0	0	0
		2020	0	0	0
12	International Journal of Advanced Research in Computer Science and Software Engineering	2016	0	1	0
		2017	0	0	0
		2018	0	0	0
		2019	0	0	0
		2020	0	0	0
13	Frontiers in Artificial Intelligence and Applications	2016	0	0	0
		2017	0	0	0
		2018	1	0	0
		2019	0	0	0
		2020	0	0	0
14	International Conference on Extending Database Technology (EDBT)	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	1	0	0
		2020	0	0	0
15	International Conference on Information Reuse and Integration for Data Science	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	0	0	0
		2020	0	0	1
16	International Journal of Learning, Teaching and Educational Research	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	00	0	1
		2020	0	0	0
17	International Journal of Recent Technology and Engineering (IJRTE)	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	1	0	0
		2020	0	0	0
18	European Journal of Operational Research	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	1	0	0
		2020	0	0	0
19	European Journal of Information Systems	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	0	0	0
		2020	0	0	1
20	International Conference on the Quality of Information and Communications Technology	2016	0	0	0
		2017	0	0	0
		2018	1	0	0
		2019	0	0	0

Table 3: Continue

		2020	0	0	0
21	MDPI Journal/ Informatics	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	0	1	0
		2020	0	0	0
22	MDPI Journal/ Publications	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	0	1	0
		2020	0	0	0
23	Journal of Digital Information Management	2016	0	0	0
		2017	0	0	0
		2018	0	1	0
		2019	0	0	0
		2020	0	0	0
24	Journal of Systems and Software	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	0	1	0
		2020	0	0	0
25	Journal of Theoretical and Applied Information Technology	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	0	1	0
		2020	0	0	0
26	The Journal of Systems & Software	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	0	0	0
		2020	0	0	1
27	Wiley Software: Evolution and process	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	1	0	0
		2020	0	0	0
28	International Journal of Recent Technology and Engineering	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	0	1	0
		2020	0	0	0
29	International Journal of Human Capital and Information Technology Professionals	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	0	0	0
		2020	0	1	0
30	Information Systems Frontiers	2016	0	0	0
		2017	0	0	0
		2018	0	0	0
		2019	1	0	0
		2020	0	0	0
Total		33	15	6	

Table 4: Quality evaluation of papers

Author	Article type	QA1	QA2	QA3	QA4	Total Score
Al-janabi and Janicki (2016)	Conference Paper	Y	N	N	P	1.
Zellal and Zaouia (2016)	Research Paper	Y	N	N	Y	2
Abdellaoui *et al*. (2016)	Research Paper	Y	N	N	Y	2
Serra and Marotta (2016)	Research Paper	Y	N	N	Y	2
Tiwari *et al*. (2017)	Research Paper	Y	P	N	P	2
Prakash and Prakash (2017)	Conference Paper	Y	N	N	Y	2
Sokolov and Turkin (2018)	Conference Paper	Y	P	N	P	2
Micic *et al*. (2017)	Conference Paper	Y	P	N	Y	2.5
Ereth (2018)	Research Paper	Y	P	N	Y	2.5
Sahoo and Premchand (2019)	Research Paper	Y	P	N	Y	2.5
Capizzi *et al*. (2019)	Research Paper	P	P	N	P	1.5
Erich *et al*. (2017)	Research Paper	P	P	N	P	1.5
Sharma (2017)	Research Paper	P	P	N	P	1.5
Trihinas *et al*. (2018)	Research Paper	P	P	N	P	1.5
Chen (2018)	Conference Paper	P	P	N	P	1.5
Snyder and Curtis (2017)	Research Paper	P	P	N	P	1.5
Zimmerer (2018)	Conference Paper	P	P	N	P	1.5
Janes *et al*., 2017	Research Paper	P	P	N	P	1.5
Jha and Khan (2018)	Research Paper	P	P	N	P	1.5
Cheriyan *et al*. (2018)	Conference Paper	Y	P	N	Y	2.5
Puonti *et al*. (2018)	Research Paper	P	P	N	P	1.5
Derakhshan *et al*. (2019)	Conference Paper	P	P	N	P	1.5
Porshnev *et al*. (2019)	Conference Paper	P	P	N	P	1.5
Deepak Raj (2019)	Research Paper	Y	P	N	Y	2.5
Lee *et al*. (2019)	Conference Paper	Y	P	N	Y	2.5
Renggli *et al*. (2019)	Research Paper	P	P	N	P	1.5
Birgersson *et al*. (2016)	Research Paper	Y	P	N	P	2
Roh *et al*. (2019)	Research Paper	P	P	N	P	1.5
Kraus *et al*. (2020)	Research Paper	Y	P	N	P	2
Chen *et al*. (2016)	Research Paper	Y	P	N	Y	2.5
Kumar *et al*. (2019)	Conference Paper	Y	P	N	Y	2.5
Nogueira *et al*. (2018)	Conference Paper	P	P	N	P	1.5
Figalist *et al*. (2019)	Research Paper	Y	P	N	P	2
Benni *et al*. (2019)	Conference Paper	Y	P	N	P	2
Chen (2018)	Conference Paper	P	P	N	P	1.5
Azeroual *et al*. (2019a)	Research Paper	Y	P	P	Y	3
Azeroual and Schöpfel (2019)	Research Paper	Y	P	P	Y	3
Azeroual *et al*. (2019b)	Research Paper	Y	Y	P	Y	3.5
Rana (2016a)	Research Paper	Y	Y	P	Y	3.5
Krishnan and Wu (2019)	Research Paper	Y	P	N	P	2
Saima Rafi (2020)	SLR	Y	Y	Y	Y	4
Meissner and Junghanns (2016)	Research Paper	Y	P	P	P	2.5
Izham Jaya (2019)	SLR	Y	P	Y	Y	3.5
Munappy *et al*. (2020)	Research Paper	Y	P	Y	Y	3.5
Tamburri *et al*. (2020)	Conference Paper	Y	P	Y	Y	3.5
Ali *et al*. (2020)	Research Paper	Y	P	N	P	2
Zarour *et al*. (2019)	Research Paper	Y	P	N	P	2
Leite *et al*. (2019)	Research Paper	Y	P	P	P	2.5
Teixeira *et al*. (2020)	SLR	P	P	N	P	1.5
Waseem *et al*. (2020)	SLR	P	P	P	P	2
Luz *et al*. (2019)	Research Paper	P	P	P	P	2
Ghantous and Gill (2019)	Research Paper	P	P	P	P	2
Koilada (2019)	Conference Paper	P	P	N	P	1.5
Hemon *et al*. (2020)	Research Paper	Y	Y	Y	Y	4
Chen (2019)	Conference Paper	P	P	N	P	1.5
Hemon-Hildgen *et al*. (2020)	Research Paper	P	N	N	P	1

Table 5: Quality average for studies by publication date

	Years				
	2016	2017	2018	2019	2020
Number of studies	8	6	13	21	6
Mean quality score	18.6	13.9	30.23	40.38	10.9

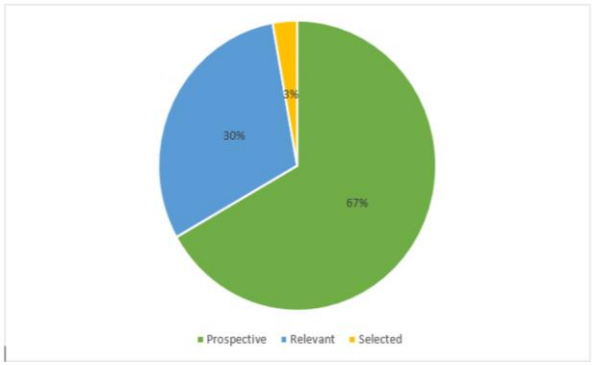

Fig. 1: Sources searched for the

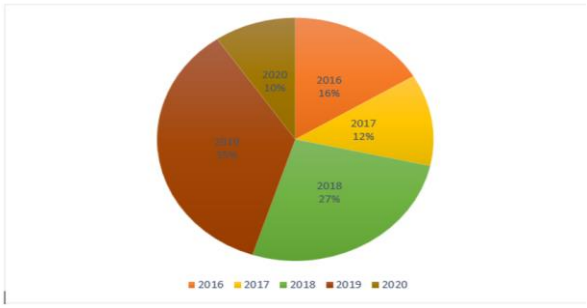

Fig. 2: Quality score of paper

Discussion

Our survey focuses on data quality literature, DevData Ops and Data Ops, all with data warehousing. It provides an exhaustive list of research papers and SLR where new techniques are proposed. The remainder of this report provides technical observations and discusses our IQs. We explore the answers to our research questions and discuss the limitations of our study. As seen in the above section of our SLR, Data Ops is an active software engineering research field. Since the advent of the Dev Ops definition, the number of publications has risen steadily.

Paper Finding Discussion

Data Ops is closely linked to the ongoing professionalization of data and analytics operations in businesses. It brings together ideas from information systems research with concepts from other fields like as agile and lean thinking, as well as current software engineering. This section presents the most related previous studies on Data Ops development and maintenance.

The research of Prabin and Anshu proposed a survey on the concept of Data Ops and how its adoption across industries gained momentum. The researchers contrasted the notions of Data Ops and Dev Ops. Then reflect on the value of Data Ops in the industrial and service sectors. Then they described the mechanism and platform of Data Ops also the data issues in the sectors of manufacturing and utilities. This study highlighted that how the usage of Data Ops can bring revolutionary improvements to a company in the analytics sector. This study clarified Data Ops can remove inefficiencies, foster cooperation and encourage reusability, lowering operating costs and speeding up time to market, by using scientific and disciplined techniques. The researchers try to summarize key components of a Data Ops platform from their point of view, they aim to break down the Data Ops process into six key phases to get the most out of Data Ops adoption. This study focus on how to build data analytics systems and did not talk about how to use Data Ops to improve data quality, whether in databases or the ETL stage (Sahoo and Premchand, 2019).

The research of Saima, Wu Yu, Muhammad, Ahmed and Abdu Gumaei, a survey aimed to critical variables that may have a detrimental effect on the method of assessing data quality in Dev Ops. The researchers used the Systematic Literature Review (SLR) method and established a total of 13 critical problems. The researchers used the Blurry TOPSIS method to prioritize the daunting analyzed variables concerning their relevance in the measurement phase of Dev Ops data quality. The findings reveal that real-time data collection, data visualization, incomplete data and other invalid data are the top-ranking issues that need to be solved on a priority basis to efficiently assess the output of heterogeneous data in Dev Ops. This study investigates a new research field in the Dev Ops domain. Using the fuzzy TOPSIS logic method, the highlighted problems were further examined in terms of their influence on Dev Ops data quality evaluation. This research focuses on the identified difficult variables that may have a detrimental influence on Dev Ops data quality evaluation procedures, on the other hand, the researchers did not talk about how to improve data quality, whether in databases or the ETL stage. They plan to undertake industrial empirical research to determine the best practices that must be followed for the Dev Ops data quality evaluation process to be effective (Saima Rafi, 2020).

The research of Aiswarya, Jan Bosc and Helena presented a direct qualitative multiple-case study and interviews with the representatives of three companies.

The researchers identified the key challenges and benefits associated with data pipeline implementation and use. They were clear about the importance of implementing data pipelines to allow traceability, fault tolerance and reduce human errors by maximizing automation to produce high-quality data. The paper's taxonomy of data pipeline challenges in Infrastructure, organizational and technological issues. The goal of this article was to look at the real-time difficulties of data pipelines and to create a taxonomy of them. In addition, it goes through the advantages of using data pipelines for creating data-intensive models. The researchers planned to expand the research with potential remedies to the problems with the data pipeline that have been identified. They did not think about how to use a data pipeline to improve data quality to use this data in the software industry (Munappy *et al*., 2020).

The research of Damian, Willem and Martin attempts to take care of knowledge issues with regards to a work market utilizing data investigation upheld by AI algorithms to enable skills localization and retrieval. A unique Data Ops model has been formulated to solve this problem by integrating data sources from administration countries and technological partners in many countries. The paper presented a model for a Data Ops intelligent analytics platform to enable a more sustainable labor market. It shows both technical feasibility and high accuracy and recalls to Data Ops intelligent analytics. On the other hand, the paper focused only on Data Ops analytics. The researchers suggested collecting hard data in the locations included in the case study to refine and put the prototype into production; and integrating the entire pipeline for abilities coordinating against existing acknowledgment to refine them to reflect the highly dynamic labor market (Tamburri *et al*., 2020).

The research of Sanjay Krishnan and Eugene Wu presented the Alpha Clean framework, which rethinks parameter adjustment in data cleaning pipelines. Users can construct data quality metrics with weighted sums of SQL aggregate queries in Alpha Clean's extensive library. Each pipelined cleaning operator adds potential transformations to a common pool in Alpha Clean's generate-then-search structure. A search algorithm arranged them into cleaning pipelines that optimize the user-defined quality metrics asynchronously, on different threads. The researchers wanted to create a system that can automatically construct and optimize data cleaning pipelines based on user-defined quality criteria. The outputs of a library of cleaning operators go into a pool of conditional assignments. The researchers present’s new framework in the data cleansing pipeline. But they did not touch on how we can use this pipeline in the data warehouse. They aimed to extend Alpha Clean towards a more flexible, visual and interactive cleaning process. They planned to integrate Alpha Clean with a data visualization system (Krishnan and Wu, 2019).

The research of Julian Ereth, a topic strategy based on a literature survey, the purpose of this study was to investigate Data Ops as a new field. It examined the collection of knowledge and offered a working meaning of Data Ops. This study examined the existing body of knowledge and offers a working definition of Data Ops also a preliminary research strategy. They began by illustrating related subjects and their systematic approach. Then they examined the preliminary findings of a qualitative investigation and develop a working definition and preliminary research methodology. This study demonstrated the wide nature of Data Ops, showing that it is a collection of principles and a way of doing things on a cultural, organizational and technological level, rather than a specific technique or instrument. However, they searched Data Ops from a theoretical point of view only. The researchers recommend doing in-depth case studies to compare traditional techniques with Data Ops-like solutions, to acquire further insights and enhance this study (Ereth, 2018).

The research of Natasha, Daniel, Felician and Esmaeil, this study specifically focuses on the data quality problem. This critical study shows that heterogeneous data sets are seldom regarded as discrete categories of data in need of a different data quality evaluation methodology during the Data Quality (DQ) assessment process. The research's ultimate goal is to provide a rigorous data quality paradigm that can be used consistently across a wide variety of heterogeneous data to provide an unambiguous indicator of data quality, independent of data specificity. Another concern raised by this analysis is the gap between conceptual frameworks and the Data Quality (DQ) evaluation criteria list method. They began to recognize the advantages of both and consider building a framework that combines the reasonable view with the physical measurement method, resulting in a more accessible Data Quality (DQ) framework for engineering domain applications. The researcher’s review of Data Quality (DQ) across different areas is done in this study to propose links between their approaches. They likewise talk about the properties of heterogeneous engineering data indexes and deteriorate their levels of heterogeneity. They give ends to the significance of data models and structures in data frameworks when creating Data Quality (DQ) assessment processes. This review highlighted the conceptual frameworks and the criteria list approach of Data Quality (DQ) assessment. On the other hand, this article dealt with data quality in theory and did not explain how to use Data Ops technology to solve the problem of data quality (Micic *et al*., 2017).

The research of Antonio, Salvatore and Manuel, This study will look into data management in Dev Ops processes, identifying relevant difficulties, challenges and potential solutions drawn from the Big Data world as well as new trends in adopting and adapting Dev Ops methods to data management. This article examined the confluence

of Dev Ops and Data Ops methods, proposing a (big) data pipeline for Dev Ops processes and toolchains additionally, it structured this pipeline according to a Data Ops process, leading to Dev Data Ops. The researchers examined the use of Data Ops in Dev Ops situations, focusing on the implementation of a Big Data pipeline and toolchain. Finally, they looked into the cutting-edge of software development, finding problems, difficulties and possible solutions. However, the study has just started to scratch the surface of such a vast subject and due to space constraints, it was impossible to cover all areas of analytics in depth (Capizzi *et al*., 2019).

The research of Otmane, Gunter and Mohammad, proposed novel data cleansing techniques for enhancing and increasing the quality of data in information systems for research. The goal was to provide potential measurements and innovative data cleaning techniques for enhancing and increasing the quality of data in information systems for research, as well as how they should be applied to research data. The researchers presented new techniques of data cleansing which can be applied to research information. On the other hand, the research did not touch the Data Ops technique (Azeroual *et al*., 2019b).

The research of Roy and Kurt, focused on proposing existing software solutions to examine and analyze the content of repositories and, in the end, to signal the Resource Description Framework's data quality and inerrability (RDF). To demonstrate its capability and confirm the usefulness of the existing RDF quality assessment tools for the Continuous Integration (CI) use case, they deployed the provided pipeline exemplarily on the Repository Hosting Services (RHS) provider GitHub3 and the Continuous Integration (CI) provider (Travis-CI). The study suggested two Continuous Integration (CI) processes for evaluating RDF data sets as well as vocabularies/ontologies for data quality and integrating them. They analyzed existing RDF quality evaluation tools and dockerized a few of them to make them easier to use in integration pipelines. However, the research focused on Dev Ops only and repurposing existing software solutions to examine and analyze (Meissner and Junghanns, 2016).

The research of Shivangi, Gagan and Kapil, this article provides an overview of the ETL Process, as well as a study of data quality issues and methods, data cleaning kinds and techniques and data cleaning types and strategies. They attempted to evaluate the data cleaning method and data error causes in this article. They also discussed the necessity of data quality, as well as the problems it poses and gave an overview of the major solution. The advantage of this article, they gave an overview of data purification in data warehouses with the use of an ETL tool. This study will aid researchers in concentrating on the many aspects of data cleaning. They explained also sources of error in data. The article briefly reviewed the data cleaning process, sources of error in data from a theoretical point of view (Rana, 2016).

The research of Daniel, Ruben, Telmo, Miguel and Joao Faustino, conducted a systematic literature review in order to identify the deciding variables that influence Dev Ops deployment, as well as the key capabilities and areas in which it grows. The objective was to have a better grasp of what Dev Ops is. By researching a subject that had previously been unexplored, this research has contributed to the academic and scientific communities. It has also enhanced the knowledge base and attempted to establish new foundations for future study. This research was a new systematized contribution to knowledge, based on the discovery of patterns that have previously been identified in the literature and so corresponds to a new degree of understanding in the approach to the issue. Professionals and practitioners will find some useful information in this study. This article was unable to gather enough data and provide a solid conclusion on particular issues such as Outcomes because Dev Ops is such a new concept (Teixeira *et al*., 2020).

In the research of Naveen and Deepika, they proposed a model called Decision Application Model (DAM). They offered a model-driven method to narrative writing that makes it more systematic and gives advice during the process. There are three layers to the strategy. The decision application model comprises (a) entities about which choices must be made and (b) inter-relationships between entities at the highest level. Entity choices are recognized at the next level. At the third level, data related to the chosen decisions are modeled and chosen to provide user stories. The DAM instance diagram is a hierarchical representation of the Application, Decision and Information levels. The focus of the DAM now moves to fill the model's ideas. Agile when extended to the Data Warehouse (DWH) does have the potential to address strategic business requirements while reducing lead-time to product delivery (Prakash and Prakash, 2017).

In the research of Aymeric, Frantz and Laetitia, this study was looking for views of job (dis)satisfaction, hazards and work circumstances of 59 individuals working in 12 agile and Dev Ops teams in the same business are investigated in this article. Dev Ops teams are defined in this article as a collection of software development and operations roles that work together to improve the quality and speed of software development and deployment processes. Operations members were included in these teams from the start of the development process, allowing development members to quickly integrate deployment limitations in production settings. They hypothesized that job happiness is linked to sharing and automation, which both encourage continual learning and experimentation but also interact and necessitate orchestration in the right work environment. The research concluded that Dev Ops delivers more work satisfaction than agile alone. This case study also revealed a risk amplification impact with Dev Ops, as well as the increased requirement to orchestrate automation and sharing,

depending on work situations. But they did not talk about DevOps with data (Hemon-Hildgen *et al*., 2020).

The Limitations of the Study

The fundamental limitations of any critical study are the partiality in Study selection and the probable inaccuracy of data extraction from the origin of the variables. The expected results of this study may be reduced Cantered on the following variables:

1. This analysis will be limited by the number of works, as the previous studies can't find several papers in different areas of the study topic.
2. Another limitation is that a few works aim to concentrate on the use of supporting DevOps tools and use them only to test their work without declaiming to the technique of DataOps.
3. To consider potential limitations that could arise from the way this study was carried out such as missing important papers and how this can be mitigated.

We have taken the following steps in designing a research strategy to remove this error and ensure accuracy and precision in our study selection:

1. The search-string-building method was viewed as a process of learning which included exploration. Subsequently, we continued our research questions to identify keywords in electronic databases for systematic study. In software engineering, search strings are language-based, so there is a risk that important studies may be missing during each search. In addition, the analysis does not take into account the alternate words used for specifications in agile processes, functionality and tasks. These words may also lead to many other studies being discovered.
2. Our research included only papers that concentrated mainly on the DevOps approach, the DataOps approach of meeting requirements, which included data quality and data scrubbing as part of it. We used studies to evaluate their validity based on inclusion and exclusion criteria to minimize alignment due to personal biases in study selection or missing any data
3. We found that there was limited information in several publications for inclusion in our research. More importantly, in the 43 papers, we found that the level of detail at which the method of study was represented varied widely throughout the studies. We also found that the authors of some studies could have selected the problems mentioned for particular reasons that are implicit and not specified in the articles

How Much Operation and Activity has been there Since 2016 with DataOps?

Total, in the sources we looked for, we found four major studies, as shown in Table 3 and Table 4. The research was classified as a study paper and a meta-analysis (Galin and Avrahami, 2006). We found two articles in open-source journals (arXiv journal and CEUR-WS Journal) and one in the International Journal of Applied Information Systems (IJAIS).

Technical Solution in our Research

We found 16 research articles, proposing and addressing information scrubbing and data consistency in the database. These articles measured, analyzed and improved the quality of data in the database area. We divided these articles into several subcategories as follows.

Data Scrubbing

Database scrubbing improved data quality by detecting and correcting low-quality data. A ton of progress has been found in this topic with most of the articles attempting to robotize the database scrubbing process and reduce user interventions. For example, Alpha Clean is a cleaning pipeline generated by a data cleaning language that can be built from the ground up building block for systems (Krishnan and Wu, 2019). That paper is tuning the pipelines of data cleaning. Another approach used a machine learning technique for automatic data correction (Al-janabi and Janicki, 2016; Derakhshan *et al*., 2019; Lee *et al*., 2019; Renggli *et al*., 2019; Birgersson *et al*., 2016). This approach learns from user feedback in the cleaning process and further refines its learning model.

Data Quality

For a long time and along these lines, data quality study has attracted researchers. The subjects of discussion are wide. We believed that classification of research topics in data goodness is major to direct researcher focus toward the least explored theme. For this reason, we found the chosen articles according to the topic discussed and we classified these research topics into three main areas including data quality impact, technical solution in the database area and technical solution in the computer science area. Most focus has been given to the technical solution in data goodness impact and technical solution in the database area (Zellal and Zaouia, 2016; Micic *et al*., 2017; Abdellaoui *et al*., 2016; Serra and Marotta, 2016; Izham Jaya, 2019). However, just five examination articles focus on the quality of data effect. It is known that the starting phase of data goodness examination incorporated the advancement of data from knowledge structure (Xiao *et al*., 2014). As we are entering the development phase of data goodness exploration, a lower number in data goodness impact can be justified. Shockingly, the quantities of surveys

in technical solutions in the database area are still low compared with other research areas.

Data Scrubbing, Data Quality and Their Impact on Data Ware House (DWH)

There is a main concern to the quality of huge data available in data warehousing technology. Discussion between an enormous amount of data and the degree of quality uncertainty of that information is still going on. The outcome demonstrated that as data volume increases, the possibility to have a data goodness problem is larger. The quality of data is central to strategies for business intelligence and data warehouse. Better data, more accurate decision choices from the process of filling data warehouse with the right data quality technologies. The data must be exact, careful, completed and consistent across data sources. The period of data quality involves terminology such as data scrubbing, data approval, data manipulation, quality data tests, data refining, data separating and tuning. It is an essential area to keep up to keep the data warehouse reliable trustworthy for business customers (Zellal and Zaouia, 2016; Serra and Marotta, 2016; Tiwari *et al*., 2017; Prakash and Prakash, 2017; Sokolov and Turkin, 2018; Rana, 2016). At the end of this Systematic Literature Review (SLR), we can bethink of a new approach in managing data scrubbing to produce data quality in an integrated database and data warehouse.

What Surveys are Being Classified in Data Ops?

Instead of basic research questions, 55 were associated with research patterns about the subject of the articles. As far as the subject field of software engineering was concerned, resulted from the research:

- Eight regarding data quality as a general-purpose
- Five linked to Data Ops as a definition and how we can use this technology.
- One linked to how to use Data Ops Pipeline in extraction and matching.
- Seventeen related to Dev Ops as a definition and how we can use this technology.
- Three linked to the relation between Continuous Deliveries (CD) and Data, data warehouse and MLOps (machine learning and operation).
- Five linked to the relation between Continuous Integration (CI) and Data, data warehouse and MLOps (machine learning and operation).
- Two linked to the relation between Data and ML Ops (machine learning and operation).
- Five linked to the relation between Dev Ops data warehouse and ML Ops (machine learning and operation).
- Eight linked to data scrubbing in the data warehouse.
- One relates to how to use Dev Ops to solve Data Quality challenges.

What Research Topics are Being Data Scrubbing by Data Ops?

From all topics, we studied and found, we did not find topics that spoke of a connection between data scrubbing by Data Ops. We tried in this Systematic Literature Review (SLR) to find methods gathering between them to enhance the ETL phase and produce data quality suitable for data warehouse and data analytic.

What are the Limitations of our Survey?

During this study, 55 data scrubbing, data quality, Dev Ops, Data Ops research articles published between 2016 and 2021 have been selected and reviewed. The selection was done dependent on the capacity of the article to answer at least one research question and that satisfied our inclusion criteria. We also consider 37 journals and six conference proceedings as article sources to include all important data quality research in this study. However, there is still a possibility to miss out on important articles especially articles that are not published in English and published in non-selected journals, conference proceedings, or magazines. We don't include quality assessment in our article selection process. This may bias the number of selected articles but we believe by collecting as many articles as we can, could help us to get a wide view of data scrubbing, data quality, Dev Ops and Data Ops research. We excluded any unrelated articles and minimize bias by imposing inclusion and exclusion criteria. However, there were generally scarcely any essential investigations in this sample. The data separated from the chosen papers was moderately objective, so we do not foresee numerous blunders in data extraction. The criteria for quality measurement are the most difficult data to extract since the DARE criteria are rather subjective. I hope that could reduce the likelihood of erroneous results.

A Systematic Literature Review (SLR) Conclusion

During this study, 55 data scrubbing, data quality, Dev Ops, Data Ops research articles published between 2016 and 2021 have been selected and reviewed. Research Papers were chosen based on their ability to give proposals in the fields of data scrubbing, data quality, Dev Ops, Data Ops, Data Ops with a data warehouse or they were able to provide solutions in those areas. Accordingly, four papers were chosen that talk about Data Ops and how to use it. We found that paper was published in 2020 to be the basis on which our survey was built. Other works were assessed and the disadvantages of each were identified using the previously established criteria. The selection was done dependent on the capacity of the article to answer at least one research question and that satisfied our inclusion criteria. We found one that relates to how to use Dev Ops to solve the Data Quality

challenge. So, we try to use Dev Ops to improve data quality in the ETL phase but we are faced with some difficulties because the Dev Ops technique is effective with a program more than a database. The interest of Data Ops organizations has led development companies to invest more in developing this field to find simple and fast ways to deal with data, as is the case in Dev Ops. A Data Ops Platform has four main software components data pipeline orchestration, quality of testing and development, automation of deployment and deployment of data science model/sandbox management (Tamburri *et al*., 2020).

Conclusion and Future Work

Data scrubbing is considered one of the most important boundaries. Data scrubbing is considered one of the most encouraging interdisciplinary developments in Information technology. This study has applied a SLR study to clarify the landscape in data scrubbing and data quality research. The growing trend in organizational use of Dev Ops activities encourages us to use Dev Ops in data scrubbing to produce clear data, as the data in the data warehouse comes from various sources and its scale is rising day by day. Nevertheless, we found some shortcomings in the Dev Ops field with data, so we resorted to the use of new technology emanating from Dev Ops called Data Ops. We supported our review with 55 published research articles. The result of this survey and our implementation indicates no one takes about the relation between data scrubbing or data quality and Dev Ops or Data Ops except five research in 2020 tried to use Dev Ops to solve data quality challenges. Therefore, we may consider ourselves the first to try to combine data cleansing with Data Ops. It is hard to build a Data Ops environment and needs a real organization change and dedication of time and resources. So, this field is still in its insufficiency and needs a lot of work to gain more accurate output to help analytics and decision-maker to take the right decision at the right time.

Based on our survey results, the field of Data Ops is still open to significant changes and developments in which the information engineering community can appreciate new and relevant solutions. We suggest some suggestions and guidance to researchers involved in designing new Data Ops with information warehouse identification techniques to help researchers. The study showed the strengths and shortcomings of the 55 academic papers analyzed derived from these criteria and recommendations. Moreover, our survey aims to improve the possibility that developers interested in analyzing and developing their applications will follow some newly suggested detection mechanisms.

Acknowledgment

We acknowledge and greatly appreciate the charge of scientific research at Helwan University for their support to complete this research. We are very grateful to the reviewers for their feedback and their invaluable help.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Abdellaoui, S., Bellatreche, L., & Nader, F. (2016, May). A quality-driven approach for building heterogeneous distributed databases: The case of data warehouses. In 2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid) (pp. 631-638). IEEE. doi.org/10.1109/CCGrid.2016.79

Ali, N., Daneth, H., & Hong, J. E. (2020). A hybrid Dev Ops process supporting software reuse: A pilot project. Journal of Software: Evolution and Process, 32(7), e2248. https://onlinelibrary.wiley.com/doi/abs/10.1002/smr.2248

Al-janabi, S., & Janicki, R. (2016, July). A density-based data cleaning approach for deduplication with data consistency and accuracy. In 2016 SAI Computing Conference (SAI) (pp. 492-501). IEEE. doi.org/10.1109/SAI.2016.7556026

Azeroual, O., & Schöpfel, J. (2019). Quality issues of CRIS data: An exploratory investigation with universities from twelve countries. Publications, 7(1), 14. doi.org/10.3390/publications7010014

Azeroual, O., Saake, G., & Abuosba, M. (2019a). Data quality measures and data cleansing for research information systems. arXiv preprint arXiv:1901.06208. https://arxiv.org/abs/1901.06208

Azeroual, O., Saake, G., & Abuosba, M. (2019b, March). ETL best practices for data quality checks in RIS databases. In Informatics (Vol. 6, No. 1, p. 10). Multidisciplinary Digital Publishing Institute. https://www.mdpi.com/2227-9709/6/1/10

Benni, B., Blay-Fornarino, M., Mosser, S., Précisio, F., & Jungbluth, G. (2019, September). When Dev Ops meets Meta-Learning: A portfolio to rule them all. In 2019 ACM/IEEE 22nd International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C) (pp. 605-612). IEEE. https://ieeexplore.ieee.org/abstract/document/8904866/

Birgersson, M., Hansson, G., & Franke, U. (2016, September). Data integration using machine learning. In 2016 IEEE 20th International Enterprise Distributed Object Computing Workshop (EDOCW) (pp. 1-10). IEEE. doi.org/10.1109/EDOCW.2016.7584357

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. Journal of systems and software, 80(4), 571-583. doi.org/10.1016/j.jss.2006.07.009

Capizzi, A., Distefano, S., & Mazzara, M. (2019, May). From Dev Ops to Dev Data Ops: Data management in Dev Ops processes. In International Workshop on Software Engineering Aspects of Continuous Development and New Paradigms of Software Production and Deployment (pp. 52-62). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-39306-9_4

Chen, B. (2019, May). Improving the software logging practices in Dev Ops. In 2019 IEEE/ACM 41st International Conference on Software Engineering: Companion Proceedings (ICSE-Companion) (pp. 194-197). IEEE. doi.org/10.1109/ICSE-Companion.2019.00080

Chen, H. M., Kazman, R., & Haziyev, S. (2016). Agile big data analytics for web-based systems: An architecture-centric approach. IEEE Transactions on Big Data, 2(3), 234-248. doi.org/10.1109/TBDATA.2016.2564982

Chen, L. (2018, April). Microservices: Architecting for continuous delivery and Dev Ops. In 2018 IEEE International conference on software architecture (ICSA) (pp. 39-397). IEEE. doi.org/10.1109/ICSA.2018.00013

Chen, L. (2018, May). Continuous delivery at scale: challenges and opportunities. In Proceedings of the 4th International Workshop on Rapid Continuous Software Engineering (pp. 42-42). doi.org/10.1145/3194760.3194764

Chen, L., Babar, M. A., & Zhang, H. (2010, April). Towards an evidence-based understanding of electronic data sources. In 14th International Conference on Evaluation and Assessment in Software Engineering (EASE) (pp. 1-4). https://www.scienceopen.com/hosted-document?doi=10.14236/ewic/EASE2010.17

Cheriyan, A., Gondkar, R. R., & Gopal, T. (2018, December). Quality Assurance Practices in Continuous Delivery-an implementation in Big Data Domain. In 2018 IEEE 8th International Advance Computing Conference (IACC) (pp. 7-13). IEEE. https://ieeexplore.ieee.org/abstract/document/8692131/

Deepak Raj, S. P. (2019). "Applying Unsupervised Machine Learning in Continuous Integration, Security and Deployment Pipeline Automation for Application Software," International Journal of Recent Technology and Engineering (IJRTE), pp, 1426-1430, 2019. doi.org/10.35940/ijrte.D7387.118419

Derakhshan, B., Mahdiraji, A. R., Rabl, T., & Markl, V. (2019, March). Continuous Deployment of Machine Learning Pipelines. In EDBT (pp. 397-408).

Dyba, T., Kitchenham, B. A., & Jorgensen, M. (2005a). Evidence-based software engineering for practitioners. IEEE software, 22(1), 58-65. doi.org/10.1109/MS.2005.6

Dyba, T., Kitchenham, B. A., & Jorgensen, M. (2005b). Evidence-based software engineering for practitioners. IEEE software, 22(1), 58-65. doi.org/10.1109/MS.2005.6

Ereth, J. (2018). Data Ops-Towards a Definition. LWDA, 2191, 104-112. http://ceur-ws.org/Vol-2191/paper13.pdf

Erich, F. M., Amrit, C., & Daneva, M. (2017). A qualitative study of Dev Ops usage in practice. Journal of software: Evolution and Process, 29(6), e1885. https://onlinelibrary.wiley.com/doi/abs/10.1002/smr.1885

Figalist, I., Biesdorf, A., Brand, C., Feld, S., & Kier Meier, M. (2019, July). Supporting the Dev Ops Feedback Loop using Unsupervised Machine Learning. In 2019 IEEE International Symposium on INnovations in Intelligent Sys Tems and Applications (INISTA) (pp. 1-6). IEEE. https://ieeexplore.ieee.org/abstract/document/8778283/

Galin, D., & Avrahami, M. (2006). Are CMM program investments beneficial? Analyzing past studies. IEEE software, 23(6), 81-87. doi.org/10.1109/MS.2006.149

Ghantous, G. B., & Gill, A. Q. (2019). An agile-Dev Ops reference architecture for teaching enterprise agile. International Journal of Learning, Teaching and Educational Research. doi.org/10.26803/ijlter.18.7.9

Hammergren, T. C. (2009). Data Warehousing for dummies. John Wiley & Sons. ISBN-10: 9780470482926.

Hemon, A., Lyonnet, B., Rowe, F., & Fitzgerald, B. (2020). From agile to Dev Ops: Smart skills and collaborations. Information Systems Frontiers, 22(4), 927-945. Hemon, A., Lyonnet, B., Rowe, F., & Fitzgerald, B. (2020). From agile to Dev Ops: Smart skills and collaborations. Information Systems Frontiers, 22(4), 927-945.

Hemon-Hildgen, A., Rowe, F., & Monnier-Senicourt, L. (2020). Orchestrating automation and sharing in DevOps teams: a revelatory case of job satisfaction factors, risk and work conditions. European Journal of Information Systems, 29(5), 474-499. https://www.tandfonline.com/doi/abs/10.1080/0960085X.2020.1782276

Inayat, I., Salim, S. S., Marczak, S., Daneva, M., & Shamshirband, S. (2015). A systematic literature review on agile requirements engineering practices and challenges. Computers in human behavior, 51, 915-929. doi.org/10.1016/j.chb.2014.10.046

Izham, J., Sidi, F., Affendey, L. S., Jabar, M. A., & Ishak, I. (2019). "Systematic review of data quality research," Journal of Theoretical and Applied Information Technology, vol. 97, no. 15 November 2019, pp. 3043-3068, 2019.

Janes, A., Lenarduzzi, V., & Stan, A. C. (2017, April). A continuous software quality monitoring approach for small and medium enterprises. In Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering Companion (pp. 97-100). https://dl.acm.org/doi/abs/10.1145/3053600.3053618

Jha, P., & Khan, R. (2018). A review paper on Dev Ops: Beginning and more to know. Int. J. Comput. Appl, 180(48), 16-20. doi.org/10.5120/ijca2018917253

Jørgensen, M. (2004). A review of studies on expert estimation of software development effort. Journal of Systems and Software, 70(1-2), 37-60. doi.org/10.1016/S0164-1212(02)00156-5

Jørgensen, M. (2007). Forecasting of software development work effort: Evidence on expert judgement and formal models. International Journal of Forecasting, 23(3), 449-462. doi.org/10.1016/j.ijforecast.2007.05.008

Khan, S. U., Niazi, M., & Ahmad, R. (2011). Factors influencing clients in the selection of offshore software outsourcing vendors: An exploratory study using a systematic literature review. Journal of systems and software, 84(4), 686-699. doi.org/10.1016/j.jss.2010.12.010

Kitchenham, B. (2004). Procedures for performing systematic reviews. Keele, UK, Keele University, 33(2004), 1-26. http://www.elizabete.com.br/rs/Tutorial_IHC_2012_files/Conceitos_RevisaoSistematica_kitchenham_2004.pdf

Kitchenham, B. A., Dyba, T., & Jorgensen, M. (2004, May). Evidence-based software engineering. In Proceedings. 26th International Conference on Software Engineering (pp. 273-281). IEEE. https://ieeexplore.ieee.org/abstract/document/1317449

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.117.471

Kitchenham, B., Mendes, E., & Travassos, G. H. (2006, April). A systematic review of cross-vs. within-company cost estimation studies. In 10th International Conference on Evaluation and Assessment in Software Engineering (EASE) 10 (pp. 1-10). doi.org/10.1109/TSE.2007.1001

Koilada, D. K. (2019, June). Business model innovation using modern DevOps. In 2019 IEEE Technology & Engineering Management Conference (TEMSCON) (pp. 1-6). IEEE. doi.org/10.1109/TEMSCON.2019.8813557

Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. European Journal of Operational Research, 281(3), 628-641. doi.org/10.1016/j.ejor.2019.09.018

Krishnan, S., & Wu, E. (2019). Alphaclean: Automatic generation of data cleaning pipelines. arXiv preprint arXiv:1904.11827. https://arxiv.org/abs/1904.11827

Kumar, R., Bansal, C., Maddila, C., Sharma, N., Martelock, S., & Bhargava, R. (2019, May). Building sankie: An ai platform for devops. In 2019 IEEE/ACM 1st International Workshop on Bots in Software Engineering (BotSE) (pp. 48-53). IEEE. https://ieeexplore.ieee.org/abstract/document/8823620/

Lee, S., Hong, S., Yi, J., Kim, T., Kim, C. J., & Yoo, S. (2019, April). Classifying false positive static checker alarms in continuous integration using convolutional neural networks. In 2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST) (pp. 391-401). IEEE. https://ieeexplore.ieee.org/abstract/document/8730202/

Leite, L., Rocha, C., Kon, F., Milojicic, D., & Meirelles, P. (2019). A survey of Dev Ops concepts and challenges. ACM Computing Surveys (CSUR), 52(6), 1-35. https://dl.acm.org/doi/abs/10.1145/3359981

Luz, W. P., Pinto, G., & Bonifácio, R. (2019). Adopting Dev Ops in the real world: A theory, a model and a case study. Journal of Systems and Software, 157, 110384. doi.org/10.1016/j.jss.2019.07.083

Mariano, D. C., Leite, C., Santos, L. H., Rocha, R. E., & de Melo-Minardi, R. C. (2017). A guide to performing systematic literature reviews in bioinformatics. arXiv preprint arXiv:1707.05813. https://arxiv.org/abs/1707.05813

Meissner, R., & Junghanns, K. (2016, September). Using Dev Ops principles to continuously monitor RDF data quality. In Proceedings of the 12th International Conference on Semantic Systems (pp. 189-192). doi.org/10.1145/2993318.2993351

Dallachiesa, M., Elmagarmid, A., Ilyas, I. F., Ouzzani, M., & Tang, N. (2013). "NADEEF: A Commodity Data Cleaning System," in SIGMOD'13 conference, New York, New York, USA, June 22–27, 2013.

Micic, N., Neagu, D., Campean, F., & Zadeh, E. H. (2017, June). Towards a data quality framework for heterogeneous data. In 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (Green Com) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (Smart Data) (pp. 155-162). IEEE. https://ieeexplore.ieee.org/abstract/document/8276745

Munappy, A. R., Bosch, J., & Olsson, H. H. (2020, November). Data Pipeline Management in Practice: Challenges and Opportunities. In International Conference on Product-Focused Software Process Improvement (pp. 168-184). Springer, Cham. doi.org/10.1007/978-3-030-64148-1_11

Nogueira, A. F., Ribeiro, J. C., Zenha-Rela, M. A., & Craske, A. (2018, September). Improving la redoute's ci/cd pipeline and devops processes by applying machine learning techniques. In 2018 11th international conference on the quality of information and communications technology (QUATIC) (pp. 282-286). IEEE. https://ieeexplore.ieee.org/abstract/document/8590203/

Zarour, M., Alhammad, N., Alenezi, M., & Alsarayrah, K. (2019). A research on Dev Ops maturity models. Int. J. Recent Technol. Eng, 8(3), 4854-4862. doi.org/10.35940/ijrte.C6888.098319

Porshnev, S., Ponomareva, O., Borodin, A., & Mirvoda, S. (2019, October). Problems and Methods for Integrating Heterogeneous Data (on Example, Metallurgical Production). In 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (Far East Con) (pp. 1-5). IEEE. doi.org/10.1109/FarEastCon.2019.8934336

Prakash, N., & Prakash, D. (2017, July). Model-driven user stories for agile data warehouse development. In 2017 IEEE 19th Conference on Business Informatics (CBI) (Vol. 1, pp. 424-433). IEEE. doi.org/10.1109/CBI.2017.67

Puonti, M., Järvi, J., & Mikkonen, T. (2018). A continuous delivery framework for business intelligence. In Information Modelling and Knowledge Bases XXIX (pp. 248-262). IOS Press. ISBN-10: 9781614998341

Rainardi, Vincent (2008). Building a data warehouse: with examples in SQL Server. John Wiley & Sons.

Renggli, C., Karlaš, B., Ding, B., Liu, F., Schawinski, K., Wu, W., & Zhang, C. (2019). Continuous integration of machine learning models with ease. ml/ci: Towards a rigorous yet practical treatment. arXiv preprint arXiv:1903.00278. https://arxiv.org/abs/1903.00278

Roh, Y., Heo, G., & Whang, S. E. (2019). A survey on data collection for machine learning: a big data-ai integration perspective. IEEE Transactions on Knowledge and Data Engineering. https://ieeexplore.ieee.org/abstract/document/8862913/

Sahoo, P. R., & Premchand, A. (2019). Data Ops in manufacturing and utilities industries. https://www.ijais.org/archives/volume12/number23/sahoo-2019-ijais-451814.pdf

Saima Rafi, W. Y. M. A. A. A. A. G. (2020). "Multicriteria based decision making of Dev Ops data quality assessment challenges using fuzzy topsis," IEEE Access, No, 17-3-2020, pp, 46958-46980, 2020. doi.org/10.1109/ACCESS.2020.2976803

Serra, F., & Marotta, A. (2016, October). Data quality in data warehouse systems: A context-based approach. In 2016 XLII Latin American Computing Conference (CLEI) (pp. 1-12). IEEE. https://ieeexplore.ieee.org/abstract/document/7833371/

Sharma, M. K. (2017). A study of SDLC to develop well engineered software. International Journal of Advanced Research in Computer Science, 8(3).

Rana, S., Negi, G. P., & Kapoor, K. (2016). "A Study over Importance of Data Cleansing in Data Warehouse," International Journal of Advanced Research in Computer Science and Software Engineering, pp. 151-157, 2016.

Snyder, B., & Curtis, B. (2017). Using analytics to guide improvement during an Agile Dev Ops transformation. IEEE Software, 35(1), 78-83. doi.org/10.1109/MS.2017.4541032

Sokolov, I., & Turkin, I. (2018, May). Resource efficient data warehouse optimization. In 2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT) (pp. 491-495). IEEE. doi.org/10.1109/DESSERT.2018.8409183

Tamburri, D. A., Van Den Heuvel, W. J., & Garriga, M. (2020, August). Data Ops for Societal Intelligence: a Data Pipeline for Labor Market Skills Extraction and Matching. In 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI) (pp. 391-394). IEEE. https://ieeexplore.ieee.org/abstract/document/9191408

Teixeira, D., Pereira, R., Henriques, T. A., Silva, M., & Faustino, J. (2020). A systematic literature review on Dev Ops capabilities and areas. International Journal of Human Capital and Information Technology Professionals (IJHCITP), 11(3), 1-22. doi.org/10.4018/IJHCITP.2020040101

Tiwari, P., Kumar, S., Mishra, A. C., Kumar, V., & Terfa, B. (2017, March). Improved performance of data warehouse. In 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 94-104). IEEE. https://ieeexplore.ieee.org/abstract/document/7975167/

Trihinas, D., Tryfonos, A., Dikaiakos, M. D., & Pallis, G. (2018). Dev Ops as a service: Pushing the boundaries of microservice adoption. IEEE Internet Computing, 22(3), 65-71. doi.org/10.1109/MIC.2018.032501519

Vinicius Lima Cruz, A. B. A. (2018). "A Dev Ops Introduction Process for Legacy Systems," in Latin American Computer Conference (CLEI), brazil, 2018.

Waseem, M., Liang, P., & Shahin, M. (2020). A systematic mapping study on microservices architecture in Dev Ops. Journal of Systems and Software, 170, 110798. doi.org/10.1016/j.jss.2020.110798

Xiao, Y., Lu, L. Y., Liu, J. S., & Zhou, Z. (2014). Knowledge diffusion path analysis of data quality literature: A main path analysis. Journal of Informetrics, 8(3), 594-605. doi.org/10.1016/j.joi.2014.05.001

Zannier, C., Melnik, G., & Maurer, F. (2006, May). On the success of empirical studies in the international conference on software engineering. In Proceedings of the 28th international conference on Software engineering (pp. 341-350). https://dl.acm.org/doi/abs/10.1145/1134285.1134333

Zellal, N., & Zaouia, A. (2016, October). A measurement model for factors influencing data quality in data warehouse. In 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt) (pp. 46-51). IEEE. doi.org/10.1109/CIST.2016.7805102

Zimmerer, P. (2018, May). Strategy for continuous testing in Dev Ops. In Proceedings of the 40th International Conference on Software Engineering: Companion Proceeedings (pp. 532-533). doi.org/10.1145/3183440.3183465